Distilled Deep Learning based Classification of Abnormal Heartbeat Using ECG Data through a Low Cost Edge Device

Morghan Hartmann, Hasan Farooq, Ali Imran
University of Oklahoma, Tulsa, USA 74135
Email: {morghan.s.hartmann-1, hasan.farooq, ali.imran}@ou.edu

***Abstract*—To meet the accuracy, latency and energy efficiency requirements of modern healthcare systems during real-time collection and analysis of health data, a distributed edge computing environment is the answer, combined with 5G speeds and modern AI techniques. Using the state-of-the-art machine learning based classification techniques plays a crucial role in creating the optimal healthcare system on the edge. This work first provides a background on the current and emerging edge computing classification techniques for healthcare applications, specifically for electrocardiogram (ECG) beat classification. After implementing these classification techniques on a Raspberry Pi-based platform we perform a comparison of the performance of these classification techniques with respect to three key performance indicators (KPI) of interest for health care applications namely accuracy, energy efficiency, and latency. Benefiting from the results of the comparative analysis presented in this work, a distilled neural network algorithm can be selected for optimal deployment and over 90% accuracy in given scenario in healthcare system depending on the specific requirements of the given scenario.**

Keywords—edge computing, deep learning, heartbeat classification, ECG

I. INTRODUCTION

Artificial intelligence (AI) has become a large research area in many industries, and one sector that could benefit the most is healthcare. A recent article by Forbes predicts that public and private sector investment in healthcare-related AI will reach $6.6 billion by 2021, which highlights the potential of AI to enable better quality of healthcare [1]. The excess of health-related data has put a strain on legacy computing used by hospitals and industry, but AI has the capacity to process and analyze this data faster and more accurately than human counterparts. Computationally complex algorithms required for healthcare data can be used in a large number of applications that help to provide life-saving medical monitoring, especially for heart failure and arrhythmia detection [2]. Deep learning specifically is an effective tool that can be utilized by the medical community, as shown by a large number of studies done in this field [3].

ECG, or Electrocardiogram, is a test that measures electrical activity of the heart. According to the American Heart Association, the test can identify parts of the heart that have been damaged, overworked, or are too large to be healthy [4]. The test is routine and harmless, as no electricity is transmitted to the body. Creating a portable ECG device can be done by utilizing energy efficient computing methods and small devices at the edge of the network. Current research in edge computing for healthcare focuses on measuring certain KPIs that are important for the progression of health services, such as response time, energy efficiency, and bandwidth cost. State-of-the-art reserach tend to focus on one of the KPIs for a certain portion of the edge computing process, so the aim of this work is to increase the number of KPIs, which provides a larger picture of best data operations techniques for classifying patient ECG data. Classification of raw data collected by health sensors is normally completed using simple or advanced algorithms, depending on the computing power of the device, and is a very common research theme in healthcare-related computing.

II. RELATED WORKS

The use of machine learning techniques for medical data classification on the edge is a common theme of today's relevant research, however, there are a few that focus specifically on ECG signal classification. For example, in [5], an Intel mote runs K-nearest neighbors (KNN) and radial basis function (RBF) algorithms for the analysis of normal and abnormal heart beats. The training data consisted of several features extracted from the raw data, including the P-Wave, RST-Wave, and T-Wave offsets. In [6], classification of ECG abnormalities is done on a PDA, which are not commonly used today. Nevertheless, the reports on classification comparisons are highly detailed, which was helpful in the selection of models for this research. The most accurate models used in [6] were the decision tree, neural network, and nearest neighbor clustering, which all had over 91% accuracy. However, the neural network had a training time of 2 hours, 10 min and the other two had times of over 5 minutes. The authors were able to classify 15 beat types from the database, in particular, the beats relating to ventricular flutter arrhythmia, which is a condition that needs medical attention in less than three minutes to avoid fatality. In [7], one-minute samples of ECG cycles, with and without arrhythmia, are analyzed using Support Vector Machine (SVM) learning. The use of this low latency technique, combined with the fog node computation in the place of cloud computing, contributes to a very low latency device (759 ms delay) that can be deployed in an IoT network. The SVM algorithm distinguishes between

978-1-7281-2999-0/19/$31.00

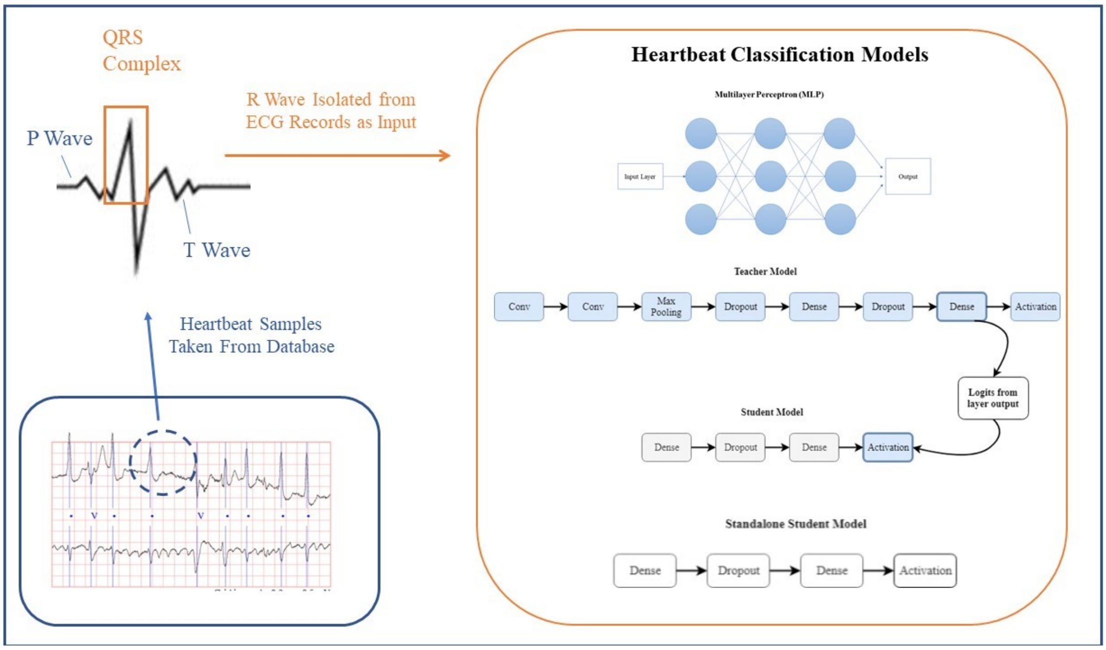

Figure 1. Derivation of ECG inputs and classification models

normal and abnormal heart rhythms based on features from the ECG with an accuracy of 93.6%. However, this work uses a different source of data, namely the “long-term ST Database” on Physiobank. The authors of [8] also use the MIT arrhythmia database and show through their work that the use of state-of-the-art deep neural networks outperforms other models in terms of accuracy and specificity.

This work adds to these previous ones by providing an analysis in memory requirements, which is especially important for small devices, and latency of runtime, which is another requirement for fast medical devices. Selection of a raw data input for the classifiers sets this project apart from previous works in ECG classification. Previous work uses a large number of features, so to test if sparse features can be used, extracted sparse raw data from the MIT database is chosen for the input to reduce the memory-related constraints as well as storage constraints. In addition to showing results for 14 beat types, we also extend to abnormal versus normal beats with a focus on distinguishing between normal beats and the most urgent beat types that need diagnosis quickly (less than three minutes). The addition of the distilled neural network in the analysis is a novel contribution to ECG edge classification, as it has not been evaluated in literature for this use. For a comprehensive diagram of this work’s setup, see Figure 1.

III. ECG Data Background

A. *ECG Data Selection*

The key parts of the ECG are the P, Q, R, and S waves (Fig. 1). The features refer to the stimulation and contraction of different parts of the heart muscle. The P-Wave is the action of the atria, or upper parts of the heart. The QRS Wave refers to the ventricles, or the lower parts of the heart contracting. The P wave signals the end of the heartbeat and represents the heart muscles resetting for another contraction sequence. P interval, QRS area, and T interval are common extracted features for the classification. Each person has a slightly different normal ECG signal, depending on gender, height, and weight, among other factors [5]. The problem with ECG data is the large amount of data that can be amassed by the sensors, especially if the number of leads is increased. Even ECG samples for small periods of time can take up megabytes of storage, which is the case for the MIT arrhythmia database, which is discussed in the next section. Each sample contains 30 minutes of two-lead data, which consists of approximately 18 MB. For a small device used on the edge of the network with limited computing and storage capabilities, this is not an acceptable amount.

B. *ECG Data and Pre-Processing*

There are several publicly available databases to use for arrhythmia classification tasks, including those on EDB, AHA, CU, and NSD. However, the most popular to use is the MIT Arrhythmia Database [9], since it has the best documentation and most beat types represented. The MIT Arrhythmia Database available on Physiobank contains 48 patient ECG records, each 30 minutes long. The heartbeats fall into five “super classes” – normal, supraventricular ectopic beats, ventricular ectopic beats, fusion beats, and unknown beats. The beats are further classified into 18 distinct types, each represented in the database records by a character. Many of the patient files contain more than one type of heartbeat. The classification task has two main parts—model formulation and prediction. The prediction of a beat into a particular class relies on a representative model, which comes from the input variables. Since these algorithms and datasets cannot exceed a certain memory usage amount, limited input observations need to be used to ensure this speed and efficiency. The input variables in this case are not the raw ECG signals, which exceed 15 MB for one patient’s data. Instead, the raw signal from the amplitude of the QRS complex is taken from each beat type in 40 different patient files. This accounts for only 160 kB once extracted, which is small enough to be run from a small device. The Pre-processing of MIT Arrhythmia Database files involved several steps. First, the files were downloaded from the Physiobank database for medical signals. The annotation files which contain the beat type labels for the duration of each patient's ECG, are separate from the raw signal files, which must be merged via an indexing program that takes the sample number from each annotation and inserts the label for the beat types in the raw signal file. Next, to select the features to be used from the data files, a filtering program was run to find the amplitude of each QRS complex, along with the raw signal information at the amplitude of both leads. These selected features were taken from multiple patient files and contained samples of 14 different beat types to analyze the robustness of the learning algorithms to classify beats for any patient.

IV. Maching Learning Models

A. Multilayer Perceptron (MLP)

Multilayer perceptron is a form of feedforward neural network, which forms a stacked regression model. The backbone of this algorithm is backpropagation, which comes from the essential knowledge of calculus [10]. MLP has three or more layers: one input layer, one output layer, and one or more hidden layers. For this work, the best performing MLP model had 3 hidden layers with 100 neurons each. The model for this scenario is the following:

$$x_n \to h_1 \to h_2 \to h_3 \to y_n \tag{5}$$

B. Deep Neural Network (DNN)

Before the distilled neural network is discussed, it is essential to know the basics of a deep neural network. This neural network is a linear stack of layers with 8 layers total. These layers consist of dense, max pooling, 1-dimensional convolution, and activation layers. The models used in this work apply weights to inputs and connect this to an output. And any input layer for the deep neural network can be represented as

$$h^{in} = M \times W + B \tag{7}$$

Where W is the weights and B represents biases from each of the neurons [11]. Neural networks with convolution layers are extremely powerful when used in many recognition tasks, and produces high accuracy results.

C. Distilled Deep Neural Network (dDNN)

As discussed above, the DNN is extremely valuable when accuracy is the goal, however, this comes at a cost of high time complexity and memory. With low latency and memory requirements to satisfy, we evaluated a smaller, more efficient model for edge computing use. This new, state-of-the-art technique aims to gather essential knowledge from one large model or a number of medium-sized models to create a "student" or generalized model that is more energy efficient than its predecessors. This small size is ideal for a small edge device. This student model distills the knowledge from a large neural network using logits, which are the inputs to the teacher's final Softmax layer [12]. Figure 1 shows the sequence of events when constructing a student model from a teacher model. The output of the teacher model's final layer is taken as an input to learn on the data at a fast rate. This in turn is used in the place of the Softmax activation layer of the student model. A complex neural network, for example, results in a high accuracy for data classification, however, it also has a high memory requirement, which is not feasible for running on a small device at the edge of the network.

V. Classification Results and Discussion

First, we analyzed the KPI that is most often used in other works—accuracy. First, the machine learning models were trained to classify 14 common beat types in an ECG. Each model was then trained to distinguish between the two beat types deemed most urgent for treatment versus normal beats. The accuracy comparisons are shown in Table 1. For comparison with the teacher-trained model, a standalone smaller model was created. This standalone model has the same structure as the trained student model, except for the activation layer which is separate from the teacher model. Thus, this entire model was trained independently from the teacher. This enables testing of the distillation compared to a similar 4 layer deep neural network that is not derived from a teacher. Although the MLP model has a relatively high accuracy that compares to the other models, it has a high loss, which cannot be overlooked when choosing an ideal classification algorithm for an edge device. The student model obtained from training on the last layer of the teacher model's logits aims to keep the accuracy at a similar level while decreasing the loss and size of the teacher. After decreasing the dataset to include only three beat types, the standalone student model achieved comparable results to the teacher-trained student model, however, the standalone had significant model loss as compared to the teacher-trained student model. Although there was a lower runtime involved with the standalone model, the accuracy and loss parameters are not as ideal as for the trained student model. Each of the classifiers chosen achieve similar accuracy results when compared, however, the teacher DNN model has a slight edge over the others. This edge comes at a cost, since it has a high runtime and the highest memory requirement out of all the classifiers.

Table 1. Accuracy Comparisons (%)

	14 beat types	3 beat types
MLP	62.5	92.1
Teacher DNN	68.3	94
Student DNN	65	89
Standalone DNN	65	90

The second metric involved in an ideal edge algorithm selection is latency, the results of which are shown in Table 2. The MLP has very low training latency, while the deep neural networks have higher latency of up to 117 seconds. The trained student model takes longer to train since some inputs from the teacher are required for the construction of the model, which increases the runtime. However, the model loss for the trained student model was slightly less than the teacher model. Because of the low loss and relatively high accuracy, this technique has the optimal traits needed for deployment in a medical edge device scenario out of all the deep learning models included.

The third metric important for deployment on edge devices is the low memory requirement, a comparison of which is shown in Table 2. Since the MLP method is not a deep learning model, it consumes less energy than the deep learning models when performed on a dataset. Although the MLP does have a higher loss, it is extremely close to the teacher model in this respect and uses almost half as less energy as the teacher model. However, if a dataset is run on a pre-trained model that can account for a variety of beat types in a variety of patients, the model might not have to be fitted to each patient's ECG. Further testing is required to test this hypothesis.

Table 2. Latency and Memory Comparisons for 14 Beat Types

	Latency (sec.)	Memory (MB)
MLP	4.89	116
Teacher DNN	117.5	244
Student DNN	115	229
Standalone DNN	107	244

VI. CONCLUSION AND FUTURE WORKS

This work provides the analysis of machine learning techniques, including a state-of-the-art method called distilled knowledge learning, which uses a large teacher model for training a smaller, edge-friendly classifier. The results of the analysis proved the usefulness of the distilled neural network, which performs with the lowest loss among the classification techniques, with only a small drop in accuracy in comparison to the large model. These smaller models are legitimate options for small devices with a larger number of features and low number of ECG beat types, which would help to decrease the loss which exists under this work's conditions. Adding more features does require more data pre-processing, which could add to the total runtime of the diagnosis program.

Based on the analysis, it is clear that using sparse data feature does limit the accuracy and precision of a system but works relatively well when the number of total classes is reduced. If more features are added, it is suggested that a smaller model such as the MLP is used for deployment in a case where low latency is the priority, above accuracy, such as an emergency diagnosis scenario. However, when all of the metrics are weighed, it is clear that the student model trained on the large deep neural network and MLP both have the best tradeoff between accuracy, loss, and runtime.

Future work would include a study on best extracted features to use in a low latency scenario. For example, comparing the total runtimes of extraction plus the learning algorithm for several different features. Adding additional data, such as heart rate or blood pressure could also add to the diagnosis aspect of a medical system. Another area of research would be to compare performance of machine learning algorithms best for specific beat types and then create specialty models to then distill into a neural network.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant Number 1559483. The statements made herein are solely the responsibility of the authors.

For details about these projects visit www.ai4networks.com

REFERENCES

[1] 'AI and Healthcare: A Giant Opportunity', *Forbes Insights*, Issue 4, 2019.

[2] O. Rajabi Shishvan, D. Zois and T. Soyata, "Machine Intelligence in Healthcare and Medical Cyber Physical Systems: A Survey," in *IEEE Access*, vol. 6, pp. 46419-46494, 2018.

[3] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, U Rajendra Acharya, Deep learning for healthcare applications based on physiological signals: A review, Computer Methods and Programs in Biomedicine, Volume 161, 2018, 1-13. 2018.

[4] "Electrocardiogram (ECG or EKG)", www.heart.org, 2018. [Online]. Available:http://www.heart.org/en/health-topics/heart-attack/diagnosing-a-heart-attack/electrocardiogram-ecg-or-ekg.

[5] Orestis Akrivopoulos, Dimitrios Amaxilatis, Athanasios Antoniou, and Ioannis Chatzigiannakis. 2017. Design and Evaluation of a Person-Centric Heart Monitoring System over Fog Computing Infrastructure. In Proceedings of the First International Workshop on Human-centered Sensing, Networking, and Systems (HumanSys'17), Rasit Eskicioglu (Ed.). ACM, New York, NY, USA, 25-30.

[6] J. Rodriguez, A. Goni and A. Illarramendi, "Real-time classification of ECGs on a PDA," in IEEE Transactions on Information Technology in Biomedicine, vol. 9, no. 1, pp. 23-34, March 2005.

[7] Iman Azimi, Arman Anzanpour, Amir M. Rahmani, Tapio Pahikkala, Marco Levorato, Pasi Liljeberg, and Nikil Dutt. 2017. HiCH: Hierarchical Fog-Assisted Computing Architecture for Healthcare IoT. ACM Trans. Embed. Comput. Syst. 16, 5s, Article 174 (September 2017), 20 pages.

[8] G. Sannino, G. De Pietro, A deep learning approach for ECG-based heartbeat classification for arrhythmia detection, Future Generation Computer Systems, Vol. 86, 446-455, 2018.

[9] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng in Med and Biol* 20(3):45-50 (May-June 2001).

[10] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*, 1st ed. Massachusetts Institute of Technology, 2012.

[11] A. Sakryukin.. "Under the Hood of Neural Networks. Part 1: Fully Connected." [online] *Towards Data Science*, 2018. Available: https://towardsdatascience.com/under-the-hood-of-neural-networks-part-1-fully-connected-5223b7f78528

[12] U. Upadhyay. "Knowledge Distillation." [online] *Towards Data* Science, 2018. Available: https://medium.com/neural-machines/knowledge-distillation-dc241d7c2322